DOCUMENT RESUME

ED 428 118 TM 029 589

AUTHOR	Sireci, Stephen G.; Fitzgerald, Cyndy; Xing, Dehui
TITLE	Adapting Credentialing Examinations for International Uses.
PUB DATE	1998-04-17
NOTE	24p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Credentials; Engineers; Factor Analysis; Foreign Countries; International Education; *Item Bias; *Licensing Examinations (Professions); Multidimensional Scaling; Occupational Tests; *Second Languages; Tables (Data); *Test Format; Test Use; Translation
IDENTIFIERS	Confirmatory Factor Analysis

ABSTRACT

Adapting credentialing examinations for international uses involves translating tests for use in multiple languages. This paper explores methods for evaluating construct equivalence and item equivalence across different language versions of a test. These methods were applied to four different language versions (English, French, German, and Japanese) of a Microsoft certification examination with samples ranging from 1,329 to 2,000 examinees per test. Principal components analysis, multidimensional scaling, and confirmatory factor analysis of these data were conducted to evaluate construct equivalence. Detection of differential item functioning across languages was conducted using the standardized p-difference index. The results indicate that these procedures provide a great deal of information useful for evaluating test and item functioning across groups. Some differences in factor and dimension loadings across groups were noted, but a common, one-factor model fit the data well. Four items were flagged for differential item functioning across all groups. Suggestions for using these methods to evaluate translated tests are provided. (Contains 8 tables, 3 figures, and 13 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 428 118

Adapting Credentialing Examinations for International Uses[1,2]

Stephen G. Sireci[3]
University of Massachusetts, Amherst

Cyndy Fitzgerald
Microsoft Corporation

Dehui Xing
University of Massachusetts, Amherst

Stephen Sireci

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029589

[1] Paper presented at the annual meeting of the American Educational Research Association, April 17, 1998, San Diego, California. This research was supported by a grant from Microsoft, Inc. to the University of Massachusetts.

[2] Laboratory of Psychometric and Evaluative Research Report No. 329. Amherst, MA: University of Massachusetts, School of Education.

[3] The authors thank Ronald Hambleton, B. Bastari, Sharon Slater, and Liane Patsula for their various and important contributions to this work.

BEST COPY AVAILABLE

Abstract

Adapting credentialing examinations for international uses involves translating tests for use in multiple languages. This paper explores methods for evaluating construct equivalence and item equivalence across different language versions of a test. These methods were applied to four different language versions of a Microsoft certification examination. Principal components analysis, multidimensional scaling, and confirmatory factor analysis of these data were conducted to evaluate construct equivalence. Detection of differential item functioning across languages was conducted using the standardized p-difference index. The results indicated that these procedures provide a great deal of information useful for evaluating test and item functioning across groups. Some differences in factor and dimension loadings across groups were noted, but a common, one-factor model fit the data well. Four items were flagged for DIF across all groups. Suggestions for using these methods to evaluate translated tests are provided.

Introduction

The information technology (IT) industry represents a relatively new and burgeoning area of certification testing. For example, IT companies such as Microsoft and Novell administer several hundred thousand certification exams each year. An important aspect of these testing programs is their international presence. On any given day, people all over the world are tested for the same credential. To meet the certification demands of today's international marketplace, tests available in multiple languages are required. Providing tests of sufficient psychometric quality in multiple languages poses new challenges for test developers and psychometricians. It is important to overcome these challenges to fulfill the fast-paced demands of the growing IT industry.

There are at least three significant challenges in IT certification examination development. First, tests need to be developed very quickly. Computer software is continually updated and the certification exams must be revised accordingly. This demand for rapid test development makes it difficult to conduct content validation and item pre-testing studies. Second, these tests need to be developed in multiple languages. Thus, the item writing phase of test development is substantially increased. Minimally, multiple steps of item translation and review are required; and often, unique items need to be constructed in the additional languages. Third, computerized testing is the rule, not the exception. Thus, test development and administration must be coordinated with computerized software[4].

Unlike many testing programs that offer tests in two major languages, such as English and Spanish, these programs offer tests in over a dozen languages. Projecting into the future, it is likely such tests will need to be made available in many more languages. Thus, similar to the situation with the Third International Mathematics and Science Study (TIMSS), which involved over 30 languages, the adaptation problem is not one of evaluating a single translation, but rather many translations. To raise awareness of the important issues involved in test translation, the International Test Commission (ITC) recently developed 22 guidelines for adapting tests from one language to another (summarized by Hambleton 1994). In considering these guidelines, Van der Vijver and Hambleton (1996) articulated three potential sources of bias affecting test translations: 1) construct bias, which describes non-equivalence of the construct measured across languages; 2) method bias, which results from problems in the administration of different language versions of a test; and 3) item bias (or "translation DIF" as termed by Berberoglu & Sireci, 1996), which may result from improper translation of an item. Consideration of these issues is further complicated for international certification programs like Novell and Microsoft due to the large number of translations that need to be made.

In evaluating different language versions of an international certification test, two conspicuous validity questions can be raised:

---

[4] To address some of the challenges faced in computerized testing, the Association of Test Publishers formed a task force to develop standards for technology-based assessment. Those interested in learning more about this task force are encouraged to contact the second author at cyndyf@microsoft.com.

1) Is the same construct measured across the different language versions of the test?

2) Is the test "equally fair" across languages?

This second question could be broken down further as:

a) Is the same level of knowledge and skill required to pass the exam consistent across languages, or is one language version "easier" than another?

b) Are there items that function differentially across the various language versions of the test?

The purposes of this paper are to present and evaluate some methodologies for evaluating construct equivalence and item equivalence across multiple language versions of a test. This paper explores these issues by applying four statistical methods to data from the Microsoft Certified Professional testing program: principal components analysis, multidimensional scaling, confirmatory factor analysis, and differential item functioning detection procedures. The next section provides a brief overview of Microsoft's certification testing program. Next, we describe the data analyzed and methods used to analyze them. Finally, we apply a series of analyses on these data to shed light on some of the comparability issues raised in adapting credentialing exams for international uses.

Overview of Microsoft's Certified Professional Program

Microsoft delivers over 600,000 exams per year in up to 75 countries. Thirty-six exams are available on a worldwide basis. In addition to English, each exam is localized (adapted for local administration) in up to 14 languages. Currently, Microsoft offers certification exams in the following languages:

• English	• Japanese	• Korean	• Simplified Chinese
• Traditional Chinese	• German	• Hungarian	• Polish
• French	• Russian	• Italian	• Spanish
• Brazilian Portuguese	• Czech	• Finnish	

Characteristics of a Microsoft certification exam

Microsoft certification exams are developed with the input of professionals in the industry and reflect how Microsoft products are used in organizations throughout the world to perform specified job functions. Microsoft certification exams typically comprise the following item types:

- Traditional multiple-choice (MC) items that measure basic knowledge and comprehension of Microsoft products and technologies.
- Scenario-based MC items that measure candidates' ability to analyze situations
- Scenario-based multiple rating items that measure candidates' ability to analyze and synthesize information and evaluate the quality of a given solution

- Simulation items that measure candidates' ability to use a simulated version of the software product

Other distinguishing features among the item types include: 1) use of graphics, tables, and other exhibits, 2) "point-and-click" items that require candidates to identify areas in a graphic, and 3) "drag and drop" items. This latter item type allows examinees to select an object (e.g., text, graphics) and move it from one location on the computer screen to another. All items on the test are scored dichotomously.

Steps in the Certification Program

To be certified as a Microsoft Systems Engineer, candidates are required to pass four operating system exams and two elective exams. To become certified as a Microsoft Solution Developer, candidates are required to pass two core technology exams and two elective exams. These exams are designed to provide a valid and reliable measure of technical proficiency and expertise. The core technology exams require candidates to demonstrate their understanding of Windows® 32-bit architecture, OLE, UI design, and Windows Open Services Architecture components. The elective exams require expertise with Microsoft development tools. When an exam is retired, the candidate generally has six months to pass an additional exam to maintain his/her certification. Otherwise, he/she will become decertified.

Adapting English-Language Tests for Use Internationally

The test adaptation process at Microsoft consists of three phases: pre-localization, localization, and post-localization. The pre-localization phase occurs concurrently with the development of the English version of the exam. During this phase, subject matter experts verify that the tasks measured in each of the items can be performed with localized versions of the software product. This phase is critical since the functionality of Microsoft software is not always consistent across languages due to constraints such as availability of specific hardware components. Items are also reviewed to determine if they meet a number of additional criteria such as the ability to localize scenarios, server names and graphics, to name a few.

The localization phase consists of the translation of exam content. Translators are provided with training and guidelines for completing these translations. Translators are instructed to translate the intent of the item instead of a word for word translation. The final phase, post-localization, involves an extensive technical review that is performed in the native country. Reviewers are provided with an electronic version of the exam so that they may view it exactly as it is going to appear to the examinee. This has become even more important to the validity of Microsoft certification exams now that these exams include simulations. Reviewers provide feedback directly into this electronic review version of this exam. Once the feedback has been verified, the exam is recompiled, published, and available worldwide. Examinees may take the exam in the language of their choice.

Method

Data

The test data analyzed come from a recent version of Microsoft's Networking Technology Server (NTS) exam, which is one of the four operating systems exams required to become a Microsoft Certified Systems Engineer. Random samples of candidates from four of the most popular language versions of the exam were selected: English (n=2,000), French (n=1,329), German (n=1,576), and Japanese (n=2,000). The NTS exam comprises 55 items measuring six global content areas: planning (5 items), installation and configuration (14 items), managing resources (10 items), connectivity (8 items), monitoring and optimization (8 items), and troubleshooting (10 items). Some descriptive statistics for the exam are presented in Table 1. The mean scores are similar across the four language versions, ranging from 36.38 (Japanese) to 39.52 (English). The standard deviations are also similar, ranging from 6.98 (English) to 7.86 (German). The KR-20 reliabilities and standard errors of measurement are also included in Table 1. The reliabilities are also similar, ranging from .82 (French) to .86 (German). The standard errors of measurement for each language version are all close to 3 points. The similarity of these descriptive statistics across the different language versions of the test is encouraging, however, they do not address questions regarding factorial or item invariance.

[Insert Table 1 Here]

Data Analysis

Evaluating construct equivalence

Principal components analyses (PCA), multidimensional scaling (MDS) analyses, and confirmatory factor analyses (CFA) were used to evaluate the structure of the examination data across the four language versions of the test. The purpose of these analyses was to discover whether the dimensionality of the examination data was consistent across language versions. If the same factor (dimensional) structure is observed across test versions, an argument can be made for construct equivalence.

Due to the potential unreliability of item-level data, the PCA analyses were conducted using both item-level data and item parcel data. The error associated with a single multiple-choice item is generally large, and so item-level factor analyses often lead to solutions with more factors than are necessary. As Dorans and Lawrence (1987) state "item level data is fraught with noise due to the unreliability of a single item, ... variation due to differences in item difficulty and examinee item responding strategies are likely to dominate item level analyses" (p. 84). Forming subgroups of items, called parcels or bundles, is a popular strategy for overcoming this problem (e.g., Cattell, 1956, Cattell & Burdsal, 1975; Dorans & Lawrence, 1987). Thirteen parcels of items were created based on the content specifications of the test and an attempt to balance the difficulty and variability of parcel scores. Parcel scores were computed by summing the scores of items comprising sub-content areas within each of the six major content areas. The thirteen parcels comprised between three to six items. The item-level PCA analyzed a matrix of inter-item tetrachoric correlations. The parcel level PCA analyzed a matrix of Pearson

correlations among the thirteen parcels. Separate matrices were derived for each language version of the test, and the PCA were conducted separately for each version.

MDS was used to evaluate the dimensionality of the data from all four language versions of the test simultaneously. The MDS analyses were performed only on the parcel data. The data for each language group was split into two random samples, and separate inter-parcel Pearson correlation matrices were computed for each sample. This procedure provided a total of eight correlation matrices for the analysis: two matrices for each language group. These matrices were fit using a weighted MDS model that allows for separate dimension weights to be derived for each proximity matrix. These dimension weights reflect the differential weighting of the dimensions necessary to best account for the correlations among the item parcels in each matrix. Thus, differences in dimensional weights across language groups would suggest differences in the dimensional structure across language versions of the test. Two proximity matrices were derived for each language group so that variation among the weights within each language could be compared with variation among the weights across language groups. The weighted MDS model used was the INDSCAL model proposed by Carroll and Chang (1970), which uses a weighted Euclidean distance formula to scale the items. The basic formula for the INDSCAL
model is:

$$d_{ijk} = \sqrt{\sum_{a=1}^{r} w_{ka} (x_{ia} - x_{ja})^2}$$

where: d_{ijk}=the Euclidean distance between points i and j for matrix k, w_{ka} is the weight for matrix k on dimension a, x_{ia}=the coordinate of point i on dimension a, and r=the dimensionality of the model. A common dimensional space (called the stimulus space) is derived for the items. The "personal" distances for each matrix are related to the group space by:

$$x_{kia} = \sqrt{w_{ka}}\, x_{ia}$$

where x_{kia} represent the coordinate for stimulus i on dimension a in the personal space for matrix k, w_{ka} represents the weight of matrix k on dimension a, and x_{ia} represents the coordinate of stimulus i on dimension a in the group space. Thus, the vector of weights $\underline{w_k}$ are used to "shrink" or "stretch" the dimensions derived from the group space to reveal the optimal dimensionality of the data for each matrix k.

MDS models fit distances to dissimilarity data, not to similarity data. Therefore, the parcel correlations were transformed to dissimilarities using the transformation suggested by Davison (1985):

$$\delta_{ij} = \sqrt{2 - 2r_{ij}}$$

where δ_{ij}=the dissimilarity between parcel i and j, and r_{ij}= the Pearson correlation between parcels i and j.

The CFA analyses were also conducted only using the parcel data. Three CFA models were fit to the data. The first model specified a single factor underlying the parcel correlations for all groups. The second model added the restriction that the factor loadings for the parcels were the same for all groups. The third model also specified that the errors associated with the factor loadings were equivalent across groups. The goodness of fit index (GFI) and root mean squared residual (RMSR) were used to evaluate data—model fit (Marsh, Balla, & MacDonald, 1988).

Evaluating item equivalence

Procedures for evaluating differential item functioning (DIF) were used to evaluate the statistical equivalence of the individual items across the four language versions of the test. DIF procedures traditionally evaluate the functioning of a single item across two groups of examinees. We use the procedures here to evaluate the functioning of different language versions of an item (i.e., English versus translated version, or two translated versions) across language groups. Three potential strategies for conducting DIF analyses across more than two different language groups were identified: 1) conducting pairwise analyses, which involves separate analyses for all possible pairs of language groups, 2) comparison of each language group to a "composite" group formed by aggregating the data for all language groups (Ellis &Kimmel, 1992), or 3) comparing each of the three "translated" groups to the English group. The pairwise approach was selected to facilitate discovery of inter-group differences.

Two factors influenced choice of the criterion used for identifying DIF items. First, was the enormous power of these analyses due to the large sample sizes involved (minimum N=1,329). Sample sizes this large will tend to yield statistically significant differences among items even when the impact of the difference is inconsequential. Second, a total of six pairwise comparisons were made for each item. Thus, using an alpha of .05, the familywise type I error rate would approach .30 for each item. Given that there are 55 items on the test, there are a total of 330 DIF comparisons to be made! Therefore, two cutoff criteria for flagging an item for DIF were established. These criteria were based on a variation of the standardized P-difference index (Dorans & Kulick,1986; Dorans & Holland, 1993) as implemented in the DICHODIF computer program (Rogers, Swaminathan, & Hambleton, 1993). The standardized P-difference (STD-P) index computed in DICHODIF is based on the differences in the proportions of examinees in the reference and focal groups, conditional on total test score, weighted by the proportion of examinees in the focal group:

$$STD-P = \frac{\sum_m w_m (E_{fm} - E_{rm})}{\sum_m w_m}$$

where w_m=the relative frequency of the reference group at score level m, and E_{fm} and E_{rm} are the proportion of examinees at score level m who answered the item correctly in the focal and reference groups, respectively. The English group served as the reference group for those comparisons in which it was involved, the French group served as the reference group for the French/German and French/Japanese comparisons, and the German group served as the reference group for the German/Japanese comparison.

Two criteria based on the STD-P index were used to flag items for DIF across languages. First, for each paired comparison, the items were rank-ordered according to the STD-P index. The top ten items with largest DIF values were considered to be potential DIF items. The second criterion was a STD-P index value of .10 or greater. This value represents a conditional p-value difference of .10, which seemed reasonable as a minimum amount of DIF that could be considered to have practical consequences. For example, if ten items exhibited DIF at this level, all of which disfavored the same group, the aggregate impact would be about one-point (one item) on the raw score scale.

Results

PCA Results

The item-level PCA results exhibited low percentages of variance accounted for by the first factor across all four groups. The variance in the item-level data accounted for by the first factor ranged from 10.4% (French) to 13.0% (German). The number of eigenvalues greater than one ranged from 16 to19 across the four groups. These results are not particularly revealing, except for confirming the expectation of a large amount of error variance present in the item-level data. Due to the large numbers of eigenvalues extracted, and the small variance accounted for by the first factors, differences among the item factor loadings across groups were not investigated. Rather, our interpretations focused on the PCAs conducted on the item parcels.

The PCA results for the thirteen item parcels were similar across the four language groups in terms of eigenvalues and percentage of variance accounted for by the first factor. In all cases, a one-factor solution fit the data well. The first factor accounted for between 31.2% (French) and 36.4% (German) of the variance among the item parcels. These results are summarized in Table 2. The eigenvalues for the first component were between 4.1 (English and French) and 4.7 (German), and the eigenvalues for the second component were all close to one. The largest proportion of variance accounted for by the second component was 9.3% for the English language sample.

[Insert Table 2 Here]

Although the separate-group PCA analyses were similar in terms of variance accounted for, there are some conspicuous differences among the factor loadings across groups. Table 3 gives the factor loading matrix for each language group. For seven of the parcels (parcels 1, 2, 4, 6, 7,12, and 13) the loadings were similar across all four groups. However, for the English data, parcels 3, 5, and 8 exhibited very small loadings on the first factor relative to the loadings for the other groups. These three parcels had very large loadings relative to the other groups on the second factor. Parcels 10 and 11 exhibited a different pattern of loadings for the Japanese group than for the other groups. For the Japanese data, parcel 10 had a large loading on factor one and parcel 11 had a loading of zero on factor one. The reverse pattern of loadings occurred on the second factor. Parcel 9 had a loading of zero on the first factor for the French data, which was small relative to the other groups. Across the four groups, the factor loadings for the French and German data appear most similar. Given these findings, it appears possible

that more than one dimension is necessary to account for the variation among the parcels across the four language groups.

[Insert Table 3 Here]

MDS Results

To evaluate the factor structure among the groups simultaneously, the data for each language group were split randomly (without replacement) and separate inter-parcel correlation matrices were derived for each sample. Two- through six-dimensional INDSCAL solutions were fit to the data. Table 4 presents descriptive fit statistics for these solutions. The STRESS index represents the square root of the normalized residual variance of the monotonic regression of the MDS distances on the transformed proximities. Thus, lower values of STRESS indicate better fit. The R^2 index reflects proportion of variance of the transformed proximities accounted for by the MDS distances. Thus, higher values of R^2 indicate better fit. Using either index, the largest improvement in fit occurs between the two- and three-dimensional solution. The three-dimensional solution accounted for 75% of the variation among the transformed parcel dissimilarities. Interpretation of the stimulus coordinates from the three- and four-dimensional solutions supported acceptance of the three-dimensional solution. The percentages of variance accounted for by the first through third dimensions were 35%, 23%, and 17%, respectively.

[Insert Table 4 Here]

The item coordinates for the three-dimensional INDSCAL solution are presented in Table 5. Sub-spaces of the solution are presented visually in Figures 1 and 2. Dimension 1 polarizes parcels 10 (monitoring/optimization) and 11 (troubleshooting). This dimension distinguishes between the proactive and reactive aspects of network technology. Interestingly, parcels 10 and 11 were the parcels that exhibited relatively different loadings in the PCA for the Japanese data. Dimension 2 separated the two parcels from the connectivity content area (parcels 7 and 8). Parcel 8 had a large negative coordinate on this dimension as did parcel 3 from the installation and configuration content area. These two parcels were two of the three parcels exhibiting relatively different PCA loadings for the English data. This dimension seems to be related to wiring issues such as installation and connectivity. Dimension 3 exhibits a more continuous ordering of the parcels with the largest distance observed between a parcel from the troubleshooting content area (13) and parcel 5 from the managing resources content area. This dimension was interpreted as ordering the parcels with respect to their relevance to managing resources. The simultaneous fitting of the data for all groups using MDS seems to reflect some of the differences observed in comparing the separate PCAs across language groups.

[Insert Table 5 Here]

[Insert Figure 1 Here]

[Insert Figure 2 Here]

The subject weights for each sample matrix are displayed in Figure 3. The German samples have the largest weights on the "troubleshooting/optimization" dimension (Dimension 1), while the English samples have the largest weights on the "wiring" dimension (Dimension 2). More equal weighting of the dimensions was required to account for the variance in the French and Japanese data. Although these differences are interesting, it is important to note that all three dimensions were relevant to the scaling of the data for all four language groups.

[Insert Figure 3 Here]

CFA Results

The results of the confirmatory factor analyses are summarized in Table 6. Inspection of the fit indices indicates that all three unidimensional models fit the data well, including the most restrictive model specifying the errors associated with the factor loadings to be equal across groups. The GFI indices for all models are above .90, and the RMSR are all below .10.

[Insert Table 6 Here]

At first glance, the CFA results seem to contradict the PCA and MDS results. Fit of a common unidimensional model to the data was not expected given the differences observed among the PCA factor loadings and the MDS dimension weights. However, the PCA and MDS results are primarily descriptive, and therefore useful for discovering differences among the groups. Although differences among the groups do seem to exist, the results of the CFA suggest these differences are not large enough to warrant different factor structures for one or more groups.

Results of DIF Analyses

Moving from evaluating factorial invariance, we now turn to analysis of item invariance. A summary of the pairwise DIF analyses is presented in Table 7. The numbers in the body of the table indicate the rank-order of the ten items with the largest STD-P DIF indices. The asterisks indicate those items with STD-P DIF indices equal to or greater than .10. The row and column totals reflect the numbers of items in the row/column with STD-P indices above .10. The pluses and minuses indicate the direction of DIF for these items, with pluses indicating larger proportion correct values for the reference group. Almost half (27 of 55) of the items were not flagged for DIF in any of the comparisons. Ten items were flagged for DIF in only one comparison. The remaining 18 items were flagged at least twice, with four items being flagged in four or more comparisons: items 37, 38, 42, and 50. The consistency of DIF observed across languages for these four items is suggestive of translation problems.

[Insert Table 7 Here]

With respect to content areas, all content areas had at least one item flagged for DIF in at least one comparison. The percentages of items flagged for DIF in at least one comparison ranged from 20% for the planning content area (one out of five items) to 75% for the monitoring and optimization content area (6 of 8 items).

Of the 330 total DIF comparisons, 62 (18.8%) had STD-P indices above .10. The largest levels of DIF were observed for those comparisons involving the Japanese items. Across the three comparisons for each language group, the mean numbers of items flagged for DIF were 9.67, 9.67, 8.67, and 13.33 for the English, French, German, and Japanese groups, respectively.

Due to the relatively larger amount of DIF observed for those comparisons involving the Japanese data, an evaluation of the DIF observed among the other groups is warranted. Table 8 presents the results of the DIF analyses for only the English, French, and German groups. Of the 165 DIF comparisons among these groups, 22 comparisons (13.3%) were flagged for DIF (down 5% in comparison to those analyses including the Japanese data). Forty-one of the 55 items were not flagged for DIF across any of the three group comparisons. Five items were flagged for DIF in two comparisons, and two (items 37 and 38) were flagged in all three comparisons. The mean number of items displaying DIF across the two comparisons for each group was 6.5, 8.5, and 7.0 for the English, French, and German groups, respectively. With respect to content areas, none of the planning items exhibited DIF across these groups. For the other five content areas, the percentage of items exhibiting DIF in at least one comparison ranged from 20% (2 of 10) for the managing resources content area to 38% (3 of 8) for the monitoring and optimization content area.

[Insert Table 8 Here]

The "practical" significance of the observed DIF can be estimated by inspecting the direction of DIF across comparisons. Given the mean levels of DIF observed for each group, if the DIF were all in the same direction (e.g., disfavoring the focal group), it would suggest an average total test score differences between 0.8 and 1.3 points across all groups. Comparison of the pluses and minuses in Table 7 suggests differences in DIF "impact" ranging from four-tenths of a point for the English/French comparison, to 1.2 points for the English-Japanese comparison. Four of the five items that were flagged for DIF in the English/German comparison favored the English group, suggesting an average impact of three-tenths of one point. Thus, the likely practical impact of the DIF observed appears small for three of the four groups. The larger impact estimated for the Japanese candidates suggests closer inspection of the translations for those items.

Inspecting the cells of Table 8 reveals some interesting parallels between the DIF results and the MDS results. For example, the DIF observed between the English and German groups was predominantly (4 of 5 items) from the managing resources and connectivity content areas. The two item parcels from the connectivity content area were strongly related to the dimension that accounted for 56% of variance in the English data, but only 7% of the variance in the German data.

Discussion

As this paper illustrates, evaluating the factorial and item invariance of the numerous tests comprising an international testing program is a complex endeavor. Obviously, the breadth of the analyses conducted in this study cannot be routinely applied to all versions of all tests in a timely manner. However, the results of these analyses can and should be used to discover common problems in translations, or important

differences among language groups. Following up the results of these analyses with content analyses to discover reasons why certain items displayed DIF, or why certain content areas seemed more relevant to the data for certain language groups, should prove instructive.

Perhaps the most important finding of this study is that PCA, MDS, and CFA, coupled with a practical approach to DIF detection, provides valuable information concerning the functioning of a test across multiple language groups. Each of the analyses conducted revealed interesting findings pointing to future research to be investigated by the test developers. For example, with respect to the NT exam, the results suggest further analysis of the four items that displayed DIF across all groups, and closer evaluation of the Japanese translation. These evaluations should help interpret the results, and suggest ways to improve future translations.

With respect to the DIF analyses, it should be remembered that the purpose of this exam was not to compare groups, or individuals across groups, with one another. Rather, this is a criterion-referenced test, and each candidate is compared to a pre-established mastery criterion. Future research should explore whether the small-to-moderate DIF observed may result in differences in the passing standard across versions of the test.

With respect to methodology, traditional approaches to the problem of evaluating construct and item equivalence were merged with newer methods based on MDS, CFA, and DIF detection methodology. MDS is useful for conducting exploratory analyses of dimensional structure when multiple groups are involved. The practical approach to DIF detection was necessitated by the multiple comparisons involved for a single item. Relying strictly on traditional statistical criteria for detecting DIF would have flagged too many items with non-meaningful differences across groups, which would also inhibit post-hoc content analyses. Finally, the results suggest that CFA is useful for determining whether observed differences across groups in factor loadings and dimension weights are sufficiently large to suggest separate scaling of the data.

In summary, conducting dimensionality and DIF studies using the procedures employed here, provides a great deal of information regarding construct and item equivalence across different language versions of a test. Follow-up content analyses of the results from these analyses should provide further guidelines for future test development.

References

Berberoglu, G. & Sireci, S. G. (1996). Evaluating translation fidelity using bilingual examinees. Laboratory of Psychometric and Evaluative Research Report No. 285. Amherst, MA: University of Massachusetts, School of Education.

Carroll, J. D., & Chang, J. J. (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika, 35, 283-319.

Cattell, R. B. (1956). Validation and intensification of the Sixteen Personality Factor Questionnaire. Journal of Clinical Psychology, 12, 205-214.

Cattell, R. B., & Burdsal, C. A., Jr. (1975). The radial parcel double factoring design: A solution to the item-vs.-parcel controversy. Multivariate Behavioral Research, 10, 165-179.

Davison, M. L., (1985). Multidimensional scaling versus components analysis of test intercorrelations. Psychological Bulletin, 97, 94-105.

Dorans, N. J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item functioning on the Scholastic Aptitude Test, Journal of Educational Measurement, 23, 355-368.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel—Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.) Differential item functioning. (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Lawrence, I. M. (1987). The internal construct validity of the SAT (Research Report 87-35). Princeton, NJ: Educational Testing Service.

Ellis, B. B. & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. Journal of Applied Psychology, 77, 177-184.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229-244.

Hambleton, R. K. & Van der Vijver (1996). Translating tests: Some practical guidelines. European Psychologist, 1 (2), 89-99.

Marsh, H., Balla, J., MacDonald, R. P. (1988). Good ness of fit indices in confirmatory factor analysis: the effects of size. Psychological Bulletin, 103, 411-423.

Rogers, H. J., Swaminathan, H. & Hambleton, R. K. (1993). DICHODIF: A FORTRAN program for DIF analysis of dichotomously scored item response data [computer program]. Amherst, MA: University of Massachusetts.

Table 1

Descriptive Statistics for NT Exams

Version	N	Mean	St. Deviation	KR-20	SEM
English	2,000	39.52	6.98	.84	2.79
French	1,329	36.50	7.05	.82	2.99
German	1,576	37.60	7.86	.86	2.94
Japanese	2,000	36.38	7.83	.85	3.03

Table 2

Summary of PCA on the 13 Parcels

Version	1st Eigenvalue	% VAF	2nd Eigenvalue	% VAF
English	4.1	32.5	1.2	9.3
French	4.1	31.2	1.0	7.7
German	4.7	36.4	0.9	7.3
Japanese	4.5	34.8	0.9	6.8

Table 3

PCA Rotated Factor Loading Matrix

		1st Factor					2nd Factor			
Content Area[a]	Parcel	English	French	German	Japanese		English	French	German	Japanese
1	1	.56	.52	.47	.51		.34	.43	.41	.32
2	2	.41	.54	.46	.56		.55	.33	.49	.30
2	3	.10	.56	.48	.58		.72	.23	.32	.29
2	4	.67	.56	.61	.53		.23	.35	.37	.47
3	5	.22	.58	.59	.46		.62	.14	.18	.33
3	6	.66	.64	.62	.51		.19	.09	.22	.29
4	7	.61	.55	.44	.44		.13	.21	.43	.49
4	8	.00	.37	.28	.48		.73	.24	.54	.22
5	9	.61	.00	.28	.56		.17	.75	.64	.16
5	10	.28	.16	-.08	.72		.39	.63	.78	-.18
6	11	.58	.59	.65	.00		-.01	-.14	.08	.79
6	12	.65	.55	.51	.58		.20	.34	.51	.36
6	13	.36	.28	.49	.32		.32	.48	.35	.50

[a]1=planning, 2=installation and configuration, 3=managing resources, 4=connectivity, 5=monitoring and optimization, and 6=troubleshooting.

Table 4

Descriptive Fit Statistics for MDS Solutions

Dimension	STRESS	R^2
2	.27	.69
3	.19	.79
4	.15	.75
5	.13	.84
6	.11	.84

Table 5

Parcel Coordinates from 3-D MDS Solution

Parcel #	Content Area	Sub-Area	Dimension 1	Dimension 2	Dimension 3
1	Planning	1.2, 1.3	.16	.32	-.83
2	Instal.&Config.	2.10, 2.4, 2.5	-.22	-.35	.60
3	Instal.&Config	2.1,2.2,2.3,2.9	.24	-1.37	.71
4	Instal.&Config	2.6, 2.7, 2.8	.29	.38	-.11
5	Manag. Resour.	3.1	-.09	-.55	1.70
6	Manag. Resour	3.3, 3.4	-.06	.74	1.25
7	Connectivity	4.1	.14	1.33	-.25
8	Connectivity	4.4	-.02	-2.31	-.20
9	Monitor/Optim.	5.1	-.63	.90	-1.72
10	Monitor/Optim.	5.2	-2.44	-.22	-.22
11	Troubleshoot	6.2, 6.4	2.47	.97	.62
12	Troubleshoot	6.3,6.5,6.6,6.7	-.31	.78	.29
13	Troubleshoot	6.8	.46	-.61	-1.85

Table 6

Summary of CFA Results

Model	GFI	RMSR
One Common Factor for all Groups	.99	.02
Equivalent Factor Loadings (Λ_x) for all Groups	.99	.03
Equivalent Errors of Factor Loadings (Θ_δ) for all Groups	.98	.03

Table 7
Summary of Pairwise DIF Analyses

Item/Content	Eng./French	Eng./German	Engl./Japan.	French/Germ.	French/Japan.	Germ./Japan	Total (*)
1 / 1							
2 / 1							
3 / 1							
4 / 1							
5 / 1			*+				1
6 / 2							
7 / 2						8*+	1
8 / 2							
9 / 2	4*-	6					1
10 / 2							
11 / 2							
12 / 2	9			10			
13 / 2							
14 / 2			6*		4*+	4*+	3
15 / 2			*+				1
16 / 2	8*+		10*+		9*-		3
17 / 2	10						
18 / 2							
19 / 2	3*-	9	*-				2
20 / 3			7*+		5*+		2
21 / 3							
22 / 3			9*+		*+		2
23 / 3						*-	1
24 / 3		2*+	4*+		7*+		3
25 / 3		5*+	3*+		*+		3
26 / 3						9*+	1
27 / 3							
28 / 3					8*+		1
29 / 3							
30 / 4				6*+			1
31 / 4							
32 / 4							
33 / 4			*+			*+	2
34 / 4							
35 / 4							
36 / 4		3*+		8*-			2
37 / 4	1*+	4*+		2*-		3*+	4
38 / 5	5*+	7	1*+	1*-	6*+	1*+	5
39 / 5			*+			7*+	2
40 / 5					10*+		1
41 / 5	6*+			4*-	3*-		3
42 / 5	7*+		5*+	7*-		5*+	4
43 / 5							
44 / 5							
45 / 5				9*-			1
46 / 6							
47 / 6							
48 / 6			8*+		2*+	6*+	3
49 / 6			*+			10*+	2
50 / 6	2*+		*-	5*-	1*-	2*-	5
51 / 6							
52 / 6		8					
53 / 6							
54 / 6							
55 / 6		1*-		3*-			2
Total (*)	**8**	**5**	**16**	**9**	**12**	**12**	**62**

Table 8
Summary of DIF Analyses with Japanese Removed

Item/Content	Eng./French	Eng./German	French/Germ.	Total (*)
1 / 1				
2 / 1				
3 / 1				
4 / 1				
5 / 1				
6 / 2				
7 / 2				
8 / 2				
9 / 2	4*-	6		1
10 / 2				
11 / 2				
12 / 2	9		10	
13 / 2				
14 / 2				
15 / 2				
16 / 2	8*+			1
17 / 2	10			
18 / 2				
19 / 2	3*-	9		1
20 / 3				
21 / 3				
22 / 3				
23 / 3				
24 / 3		2*+		1
25 / 3		5*+		1
26 / 3				
27 / 3				
28 / 3				
29 / 3				
30 / 4			6*+	1
31 / 4				
32 / 4				
33 / 4				
34 / 4				
35 / 4				
36 / 4		3*+	8*-	2
37 / 4	1*+	4*+	2*-	3
38 / 5	5*+	7	1*-	3
39 / 5				
40 / 5				
41 / 5	6*+		4*-	2
42 / 5	7*+		7*-	2
43 / 5				
44 / 5				
45 / 5			9*-	1
46 / 6				
47 / 6				
48 / 6				
49 / 6				
50 / 6	2*+		5*-	2
51 / 6				
52 / 6		8		
53 / 6				
54 / 6				
55 / 6		1*-	3*-	2
Total (*)	**8**	**5**	**9**	**22**

Figure 1

MDS 2-D Stimulus Subspace

(Dimension 1 versus Dimension 2)

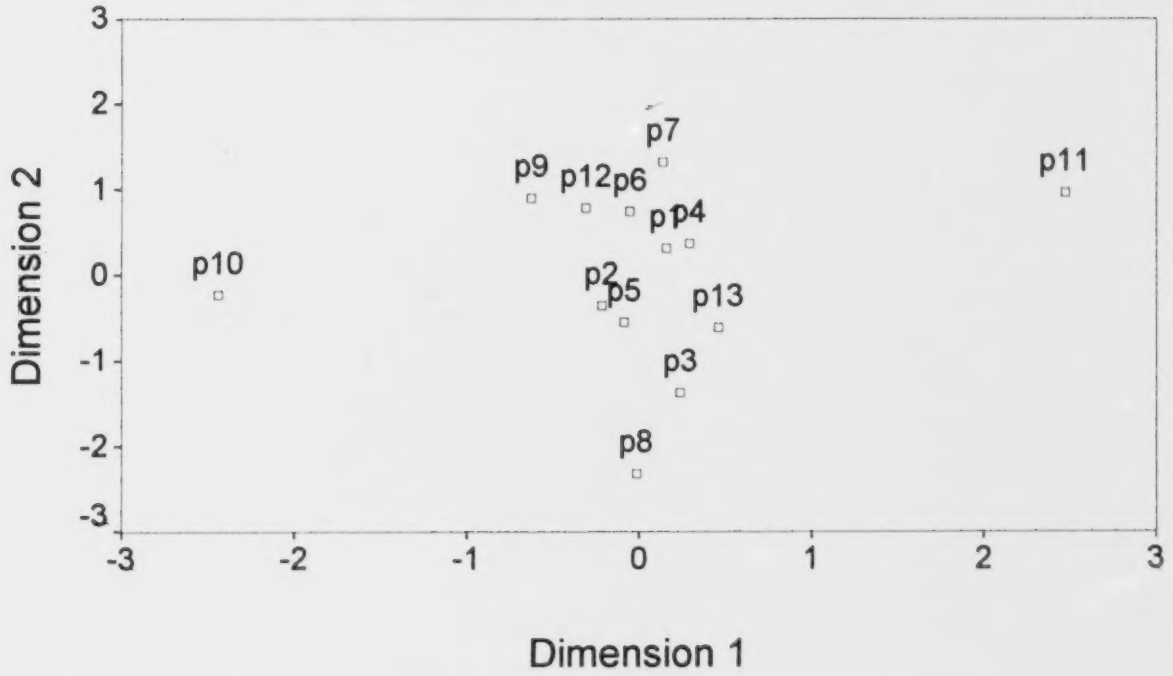

Figure 2

MDS 2-D Stimulus Subspace

(Dimenison 3 versus Dimension 2)

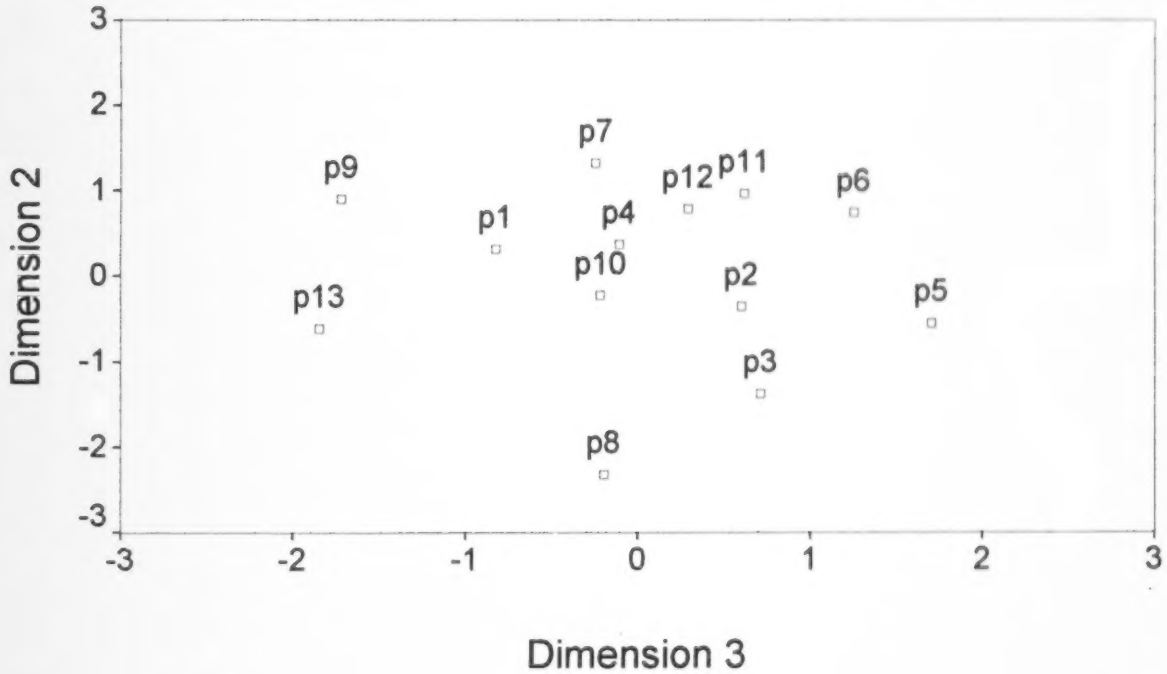

Figure 3

MDS Weight Space for NTS Parcels